(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 November 2002 (28.11.2002)

PCT

(10) International Publication Number
WO 02/094992 A2

(51) International Patent Classification[7]: C12N

(21) International Application Number: PCT/US02/15968

(22) International Filing Date: 20 May 2002 (20.05.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/291,871 18 May 2001 (18.05.2001) US

(71) Applicant (*for all designated States except US*): **RIGEL PHARMACEUTICALS, INCORPORATED** [US/US]; 240 East Grand Avenue, South San Francisco, CA 94080 (US).

(72) Inventor; and
(75) Inventor/Applicant (*for US only*): **PEELLE, Beau** [US/US]; 28 Windsor Road, Somerville, MA 02144-3135 (US).

(74) Agents: **PARENT, Annette, S.** et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, Eighth Floor, San Francisco, CA 94111 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

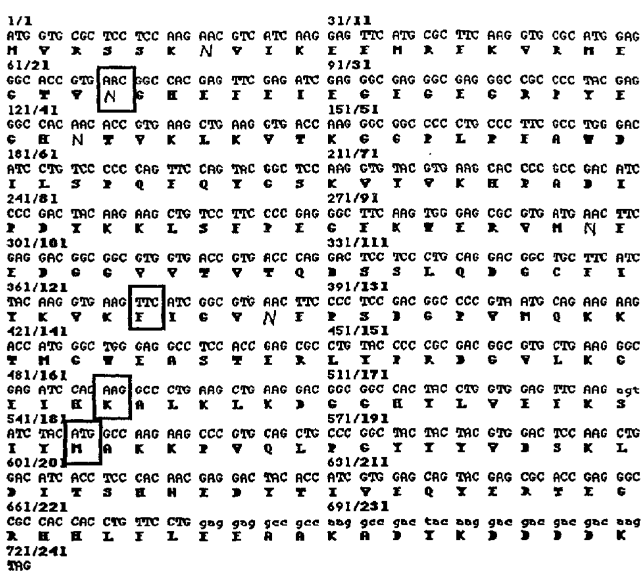
(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: DIRECTED EVOLUTION OF PROTEIN IN MAMMALIAN CELLS

Location of point mutations which enhance fluorescence of DsRED in mammalain cells

DNA and translated protein sequence of codon optimized wt DsRED protein



List of confirmed Mutations from screen:
N24S, N24R, N24H (orange box)
F125L, and F125V (blue box)
K164M (green box)
M183K (red box)

(57) Abstract: The present invention relates to directed protein evolution in mammalian cells and improved mutants of Discosoma sp. red fluorescent proteins.

WO 02/094992 A2

European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

DIRECTED EVOLUTION OF PROTEIN IN MAMMALIAN CELLS

CROSS-REFERENCES TO RELATED APPLICATIONS

The present application claims priority to USSN 60/291,871, filed May 18, 2001, herein incorporated by reference in its entirety.

STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH AND DEVELOPMENT

Not applicable.

FIELD OF THE INVENTION

The present invention relates to directed protein evolution in mammalian cells and improved mutants of *Discosoma* sp. red fluorescent proteins.

BACKGROUND OF THE INVENTION

Red fluorescent protein has been isolated from a *Discosoma* sp. and sequenced (*see, e.g.,* Matz *et al., Nature Biotech.* 17:969-973 (1999), Gross *et al., Proc. Nat'l Acad. Sci. USA* 97:11990-11995 (2000)). A variant with humanized codons has also been engineered (Clontech, "DsRED™"). The crystal structure of red fluorescent protein has been elucidated, which demonstrated that red fluorescent protein is a tetrameric protein (Wall *et al., Nat. Struc. Biol.* 7:1089 (2000); Yarbrough *et al., Proc. Nat'l Acad. Sci USA* 16:462-467 (2000)).

Red fluorescent protein (RFP) and DsRED, as well as other fluorescent proteins such as YFP, or GFP from *Aequorea victoria, Renilla reniformis, Renilla muelleri,* and *Ptilosarcus gurneyi*, are useful are reporter molecules for a variety of bioassays, including those that use FACS as a selection mechanism (*see, e.g.*, Tsein, *Nature Biotechnology* 17:956 (1999); Tsein, *Ann. Rev. Biochem.* 6:509-544 (1998); Heim *et al., Nature* 373:663-664 (1995); Heim *et al., Proc. Nat'l Acad. Sci. USA* 91:1250 (1994); Prasher *et al., Gene* 111:229 (1992); Prasher *et al., Trends in Genetics* 11:320 (1995); Chalfie *et al., Science* 263:802 (1994); and WO 95/21191). However, brighter, faster folding, and higher expressing variants would be useful.

Such variants can be made, e.g., using methods of gene shuffling and mutagenesis (*see, e.g.*, US Patent 5,811,238; WO 00/73433; WO 00/22115; WO 99/41369;

WO 01/04287; WO 00/46344; WO 99/45143, WO 99/41368; and Ichiro *et al., Protein Science* 8:731-740 (1999)). However, the use of such methods for production of variant proteins such as *Discosoma* red fluorescent protein variants is not always successful (*see, e.g.*, Baird *et al., Proc. Nat'l Acad. Sci. USA* 97:11984-11989 (2000)). Novel methods of making such variants would therefore be useful.

SUMMARY OF THE INVENTION

The present invention therefore provides variants of *Discosoma* red fluorescent protein that have been generated using directed molecular evolution in mammalian cells. The variants of the invention have greatly improved brightness, expression, and/or folding kinetics as compared to wild type or a codon optimized variant. The present invention also provides novel methods of directed protein evolution in mammalian cells using retroviral gene transfer and FACS sorting. Such methods can be used to provide improved variants of fluorescent proteins such as Discosoma red fluorescent protein and fluorescent proteins from other sources, such as *Aequorea victoria, Renilla reniformis, Renilla muelleri,* and *Ptilosarcus gurneyi.*

In one aspect, the present invention provides an isolated *Discosoma* red fluorescent protein, the protein comprising an amino acid sequence as shown in Figure 1 with one or more point mutations at an amino acid position selected from the group consisting of N24, F125, K164, and M183.

In one embodiment, the protein comprises two, three, or four point mutations at an amino acid position selected from the group consisting of N24, F125, K164, and M183.

In one embodiment, the point mutation at amino acid position N24 is a serine, arginine, or histidine substitution. In another embodiment, the point mutation at amino acid position F125 is a leucine or valine substitution. In another embodiment, the point mutation at amino acid position K164 is a methionine substitution. In another embodiment, the point mutation at amino acid position M183 is a lysine or threonine substitution.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine or valine substitution at amino acid position F125 and a lysine substitution at amino acid position M183. In another embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125 and a lysine substitution at amino acid position M183. In another embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a valine substitution at amino acid position F125 and a lysine substitution at amino acid position M183.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine or valine substitution at amino acid position F125 and a serine, arginine, or histidine substitution at amino acid position N24. In another embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125 and a serine substitution at amino acid position N24.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine or valine substitution at amino acid position F125, a serine, arginine, or histidine substitution at amino acid position N24, and a lysine substitution at amino acid position M183. In another embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125, a serine substitution at amino acid position N24, and a lysine substitution at amino acid position M183.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a methionine substitution at amino acid position K164.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125.

In one embodiment, the protein further comprises one or more point mutations at an amino acid position selected from the group consisting of K93, R19, K139, E150, and D171. In another embodiment, the point mutation at amino acid position K93 is an arginine substitution. In another embodiment, the point mutation at amino acid position R19 is a histidine substitution. In another embodiment, the point mutation at amino acid position E150 is an aspartic acid substitution. In another embodiment, the point mutation at amino acid position D171 is a glycine substitution.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125, a serine substitution at amino acid position N24, a lysine substitution at amino acid position M183, and a histidine substitution at amino acid position R19.

In one embodiment, the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125, a aspartic acid substitution at amino acid position E150, and a glycine substitution at amino acid position D171.

In one aspect, the present invention provides a *Discosoma* red fluorescent protein that is a fusion protein.

In another aspect, the present invention provides a nucleic acid encoding the *Discosoma* red fluorescent protein of the invention. In one embodiment, the nucleic acid is codon-optimized for mammalian expression. In another embodiment, the nucleic acid encodes a fusion protein.

In another aspect, the present invention provides a vector comprising a nucleic acid encoding the *Discosoma* red fluorescent protein of the invention. In one embodiment, the vector is a retroviral vector.

In another aspect, the present invention provides a host cell comprising the vector of the invention.

In another aspect, the present invention provides a retroviral cDNA expression library comprising a nucleic acid encoding the *Discosoma* red fluorescent protein.

In another aspect, the present invention provides a method of making a protein variant, the method comprising the steps of: (i) mutating a selected nucleotide sequence encoding a fluorescent protein; (ii) cloning the mutated sequences into an expression vector; (iii) transfecting mammalian cells with the expression vector; and (iv) identifying the variants.

In one embodiment, the protein is a fluorescent protein and variants are identified by FACS analysis. In another embodiment, the selected nucleotide sequence encodes a fluorescent protein from *Discosoma "red" sp.*, *Aequorea victoria*, *Renilla reniformis*, *Renilla muelleri*, or *Ptilosarcus gurneyi.*

In one embodiment, the selected nucleotide sequence is mutated using error prone PCR.

In another embodiment, the expression vector is a retroviral expression vector.

In another aspect, the present invention provides a method of making a fluorescent protein variant, the method comprising the steps of: (i) mutating by error prone PCR a selected nucleotide sequence encoding a fluorescent protein; (ii) cloning the mutated sequences into a retroviral expression vector; (iii) transfecting mammalian cells with the expression vector; and (iv) selecting variants using FACS analysis.

In one embodiment, the selected nucleotide sequence encodes a fluorescent protein from *Discosoma "red" sp.*, *Aequorea victoria*, *Renilla reniformis*, *Renilla muelleri*, or *Ptilosarcus gurneyi.*

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 provides the amino acid and nucleotide sequence of a mammalian codon optimized *Discosoma* red fluorescent protein. This figure also indicates preferred point mutations in the amino acid sequence for variants.

Figure 2 provides examples of brighter *Discosoma* red fluorescent protein variants.

Figure 3 provides a list of mutated Discosoma red fluorescent proteins isolated using the mammalian directed evolution methods of the invention.

Figure 4 shows excitation and emission spectra of certain mutants of the invention.

Figure 5 shows a diagram of methods for directed evolution of proteins in mammalian cells.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides variants of *Discosoma* red fluorescent protein, which have enhanced brightness, expression, and/or folding kinetics. These improved characteristics are useful for functional screens as a reporter for gene transcription (e.g., as a fusion protein), for target characterization and localization of fusion proteins, and for scaffolds for protein and peptide libraries. For example, variants of the invention can be cloned into expression vectors that are used to express cDNA or random peptide libraries. The variant is positioned in the vector such that it forms a fusion protein with the expressed cDNA or peptide. The cDNA library can comprise sense, antisense, full length, and truncated cDNAs. The peptide library is encoded by nucleic acids. cDNA libraries are made from any suitable RNA source. Libraries encoding random peptides are made according to techniques well known to those of skill in the art (*see, e.g.*, U.S. Patent No. 6,153,380, 6,114,111, and 6,180,343). Any suitable vector can be used for the cDNA and peptide libraries, including, e.g., retroviral vectors. The *Discosoma* variant can thus be used as a selectable marker.

Red fluorescent protein is generally useful for screens employing FACS assays. Red fluorescent protein is also useful in screens for reporter gene transcription, fusion protein localization, yeast two hybrid experiments, immunoprecipitation and proteomics, increased affinity of receptors for fluorescently labeled ligands, proteins which increase the expression level of a second protein, altered immunogenicity for fluorescently labeled antibodies, changes in cell shape and size, changes in proton pump activity, relative

DNA content in cell cycle and apoptosis, cellular localization and changes in metabolic rates of calcium flux, cell division, mitochondrial activity, pH, and free radical production. Such assays are useful for identifying proteins involved in the cell cycle, cellular proliferation, lymphocyte activation, ubiquitination pathways, cancer, mast cell degranulation, viral replication and translation (e.g., HCV) and angiogenesis. In addition to red fluorescent protein, such screens can also use one or more additional fluorescent protein, such as *Aequorea victoria* GFP, *Zoanthus* YFP and GFP, *Aneomonia* CFP, *Clavularia* CFP, *D. striata* CFP, *Renilla muelleri* GFP, *Renilla reniformis* GFP, and *Ptilosarcus gurneyi* GFP, and variants thereof.

In the present invention, novel methods of directed protein evolution were used to obtain improved variants of red fluorescent protein, as well as other proteins, including other fluorescent proteins as described above. In the methods of the invention, error prone PCR is used to randomly mutagenize a nucleic acid sequence encoding a protein of interest (*see, e.g.*, Leung *et al., Techniques* 1:11-15 (1989); Calwell & Joyce, *PCR Methods and Applications*, 2:28-33 (1992); and Gramm *et al., Proc. Nat'l Acad. Sci. USA* 89:3576-3580 (1992)). The inherently low fidelity of Taq polymerase or other thermostable polymerases can be further decreased by the addition of Mn+, increasing the Mg2+ concentration, and using unequal dNTP concentrations. A preferred method of EP-PCR is described in Calwell & Joyce, *PCR Methods and Applications*, 2:28-33 (1992) and in *Current Protocols, supra*. Alternatively, other well know mutagenesis methods such as gene shuffling could be employed (*see, e.g.*, US Patent 5, 811,238, WO 99/41369, WO 99/41368, and WO 00/46344). The library of variant nucleic acids is then transferred to mammalian cells (e.g., Jurkat, A549, Phoenix A, or BJAB) using retroviral vectors. Variants are detected by any suitable assay, e.g., in the case a fluorescent protein, by FACS. Clones of interest are then rescued and isolated. As described in Figure 1, this technique was used to identify four preferred sites of point mutations (amino acid substitutions) that lead to red fluorescent proteins with enhanced brightness, altered emission, higher expression, and/or enhanced folding kinetics (3° or 4° structure).

Definitions

The term "point mutation" refers to a deletion, addition, or substitution at a designed amino acid position in an amino acid or nucleotide sequence. Preferably, the term refers to an amino acid substitution.

"*Discosoma* red fluorescent protein" refers to a wild-type protein isolated from Discosoma species "red" (described and sequenced in Matz *et al., Nature Biotechnology* 17:969-973 (1999)), as well as a mammalian codon-optimized variant shown in Figure 1.

"Nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in single- or double-stranded form, or complements thereof. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid, and which are metabolized in a manner similar to the reference nucleotides. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, peptide-nucleic acids (PNAs). Nucleic acids also include complementary nucleic acids.

Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (e.g., degenerate codon substitutions) and complementary sequences, as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al., Nucleic Acid Res.* 19:5081 (1991); Ohtsuka *et al., J. Biol. Chem.* 260:2605-2608 (1985); Rossolini *et al., Mol. Cell. Probes* 8:91-98 (1994)). The term nucleic acid is used interchangeably with gene, cDNA, mRNA, oligonucleotide, and polynucleotide.

A particular nucleic acid sequence also implicitly encompasses "splice variants." Similarly, a particular protein encoded by a nucleic acid implicitly encompasses any protein encoded by a splice variant of that nucleic acid. "Splice variants," as the name suggests, are products of alternative splicing of a gene. After transcription, an initial nucleic acid transcript may be spliced such that different (alternate) nucleic acid splice products encode different polypeptides. Mechanisms for the production of splice variants vary, but include alternate splicing of exons. Alternate polypeptides derived from the same nucleic acid by read-through transcription are also encompassed by this definition. Any products of a splicing reaction, including recombinant forms of the splice products, are included in this definition.

The terms "polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding

naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymer.

The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, e.g., hydroxyproline, γ-carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, i.e., an α carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, e.g., homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs have modified R groups (e.g., norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refers to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid.

Amino acids may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

"Conservatively modified variants" applies to both amino acid and nucleic acid sequences. With respect to particular nucleic acid sequences, conservatively modified variants refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given protein. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation of the nucleic acid. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine, and TGG, which is ordinarily the only codon for tryptophan) can be modified to yield a functionally identical molecule. Accordingly, each silent variation of a nucleic acid which encodes a

polypeptide is implicit in each described sequence with respect to the expression product, but not with respect to actual probe sequences.

As to amino acid sequences, one of skill will recognize that individual substitutions, deletions or additions to a nucleic acid, peptide, polypeptide, or protein sequence which alters, adds or deletes a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration results in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. Such conservatively modified variants are in addition to and do not exclude polymorphic variants, interspecies homologs, and alleles of the invention.

The following eight groups each contain amino acids that are conservative substitutions for one another: 1) Alanine (A), Glycine (G); 2) Aspartic acid (D), Glutamic acid (E); 3) Asparagine (N), Glutamine (Q); 4) Arginine (R), Lysine (K); 5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); 6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W); 7) Serine (S), Threonine (T); and 8) Cysteine (C), Methionine (M) (*see, e.g.*, Creighton, *Proteins* (1984)).

The term "recombinant" when used with reference, *e.g.*, to a cell, or nucleic acid, protein, or vector, indicates that the cell, nucleic acid, protein or vector, has been modified by the introduction of a heterologous nucleic acid or protein or the alteration of a native nucleic acid or protein, or that the cell is derived from a cell so modified. Thus, for example, recombinant cells express genes that are not found within the native (non-recombinant) form of the cell or express native genes that are otherwise abnormally expressed, under expressed or not expressed at all.

The term "heterologous" when used with reference to portions of a nucleic acid indicates that the nucleic acid comprises two or more subsequences that are not found in the same relationship to each other in nature. For instance, the nucleic acid is typically recombinantly produced, having two or more sequences from unrelated genes arranged to make a new functional nucleic acid, *e.g.*, a promoter from one source and a coding region from another source. Similarly, a heterologous protein indicates that the protein comprises two or more subsequences that are not found in the same relationship to each other in nature (*e.g.*, a fusion protein).

A "fluorescent" label may be detected by exciting the fluorochrome with the appropriate wavelength of light and detecting the resulting fluorescence. The fluorescence may be detected visually, or by the use of electronic detectors such as charge coupled devices

(CCDs) or photomultipliers and the like. Similarly, enzymatic labels may be detected by providing the appropriate substrates for the enzyme and detecting the resulting reaction product. FACS analysis is a preferred method of detection when the label is in a cell.

EXAMPLES

The following example is provided by way of illustration only and not by way of limitation. Those of skill in the art will readily recognize a variety of noncritical parameters that could be changed or modified to yield essentially similar results.

Example 1: Error prone PCR and directed evolution of *Discosoma* red species fluorescent proteins in mammalian cells

To mutagenize *Discosoma* red fluorescent protein, a mammalian codon optimized variant (see Figure 1) was cloned with a flag tag and mutagenized using error prone PCR according to methods known to those of skill in the art (*see, e.g., Current Protocols in Molecular Biology*, volume 1, unit 8.3 (Ausubel *et al.*, eds, 1994); Saiki *et al., Science* 239:487 (1988); Leung *et al., Technique* 1:11-15 (1989); Caldwell & Joyce, *PCR Methods and Applications* 2:28-33 (1992); and Gramm *et al., Proc. Nat'l Acad. Sci. USA* 89:3576-3580 (1992)).

The resulting library of mutagenized sequences was cloned into a retroviral vector expression library using RT-PCR and the retroviral library was used to infect human cells (BJAB cells). The cells were sorted for brighter fluorescence, higher expression, or shifted emission. Selected clones were isolated using RT-PCR, and sub-libraries were constructed and selected further with FACS (see Figure 5). Single cell clones were isolated and sequenced.

Figure 2 lists some of the brighter mutants identified in the screen (note: amino acid sequences are off by one from the sequence numbering described in Figure 1 as the methionine was counted as zero for the purposes of the numbering in Figure 2). Figure 1 lists certain preferred mutations at amino acid positions N24, F125, K164 and M183, e.g., N24S/R/H; K125L/V; K164M; and M183K. These mutations can exist in the mutated variants alone or in any combination of one , two , three, or four, or optionally with additional point mutations at K93, R19, K139, E150, and D171, e.g., K93R, R19H, E150D, and D171G.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to one of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

WHAT IS CLAIMED IS:

1. An isolated *Discosoma* red fluorescent protein, the protein comprising an amino acid sequence as shown in Figure 1 with one or more point mutations at an amino acid position selected from the group consisting of N24, F125, K164, and M183.

2. The protein of claim 1, wherein the protein comprises two, three, or four point mutations at an amino acid position selected from the group consisting of N24, F125, K164, and M183.

3. The protein of claim 1, wherein the point mutation at amino acid position N24 is a serine, arginine, or histidine substitution.

4. The protein of claim 1, wherein the point mutation at amino acid position F125 is a leucine or valine substitution.

5. The protein of claim 1, wherein the point mutation at amino acid position K164 is a methionine substitution.

6. The protein of claim 1, wherein the point mutation at amino acid position M183 is a lysine or threonine substitution.

7. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine or valine substitution at amino acid position F125 and a lysine substitution at amino acid position M183.

8. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125 and a lysine substitution at amino acid position M183.

9. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a valine substitution at amino acid position F125 and a lysine substitution at amino acid position M183.

10. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine or valine substitution at amino acid position F125 and a serine, arginine, or histidine substitution at amino acid position N24.

11. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125 and a serine substitution at amino acid position N24.

12. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine or valine substitution at amino acid position F125, a serine, arginine, or histidine substitution at amino acid position N24, and a lysine substitution at amino acid position M183.

13. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125, a serine substitution at amino acid position N24, and a lysine substitution at amino acid position M183.

14. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a methionine substitution at amino acid position K164.

15. The protein of claim 1, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125.

16. The protein of claim 1, further comprising one or more point mutations at an amino acid position selected from the group consisting of K93, R19, K139, E150, and D171.

17. The protein of claim 16, wherein the point mutation at amino acid position K93 is an arginine substitution.

18. The protein of claim 16, wherein the point mutation at amino acid position R19 is a histidine substitution.

19. The protein of claim 16, wherein the point mutation at amino acid position E150 is an aspartic acid substitution.

20. The protein of claim 16, wherein the point mutation at amino acid position D171 is a glycine substitution.

21. The protein of claim 18, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125, a serine substitution at amino acid position N24, a lysine substitution at amino acid position M183, and a histidine substitution at amino acid position R19.

22. The protein of claim 19, wherein the protein comprises an amino acid sequence as shown in Figure 1 with a leucine substitution at amino acid position F125, a aspartic acid substitution at amino acid position E150, and a glycine substitution at amino acid position D171.

23. A fusion protein comprising the protein of claim 1,

24. An isolated nucleic acid encoding the protein of claim 1.

25. The nucleic acid of claim 24, wherein the nucleic acid is codon-optimized for mammalian expression.

26. The nucleic acid of claim 24, wherein the nucleic acid encodes a fusion protein.

27. A vector comprising the nucleic acid of claim 24.

28. The vector of claim 27, wherein the vector is a retroviral vector.

29. A host cell comprising the vector of claim 27.

30. A retroviral cDNA expression library comprising the nucleic acid of claim 24.

31. A retroviral cDNA expression library encoding the protein of claim 1.

32. A method of making a fluorescent protein variant, the method comprising the steps of:

(i) mutating a selected nucleotide sequence encoding a fluorescent protein;
(ii) cloning the mutated sequences into an expression vector;
(iii) transfecting mammalian cells with the expression vector; and
(iv) selecting variants using FACS analysis.

33. The method of claim 32, wherein the selected nucleotide sequence is mutated using error prone PCR.

34. The method of claim 32, wherein the expression vector is a retroviral expression vector.

35. The method of claim 32, wherein the selected nucleotide sequence encodes a fluorescent protein from *Discosoma "red" sp.*, *Aequorea victoria*, *Renilla reniformis*, *Renilla muelleri*, or *Ptilosarcus gurneyi.*

36. A method of making a fluorescent protein variant, the method comprising the steps of:

(i) mutating by error prone PCR a selected nucleotide sequence encoding a fluorescent protein;

(ii) cloning the mutated sequences into a retroviral expression vector;

(iii) transfecting mammalian cells with the expression vector; and

(iv) selecting variants using FACS analysis.

37. The method of claim 36, wherein the selected nucleotide sequence encodes a fluorescent protein from *Discosoma "red" sp.*, *Aequorea victoria*, *Renilla reniformis*, *Renilla muelleri*, or *Ptilosarcus gurneyi.*

Figure 1: Location of point mutations which enhance fluorescence of DsRED in mammalain cells

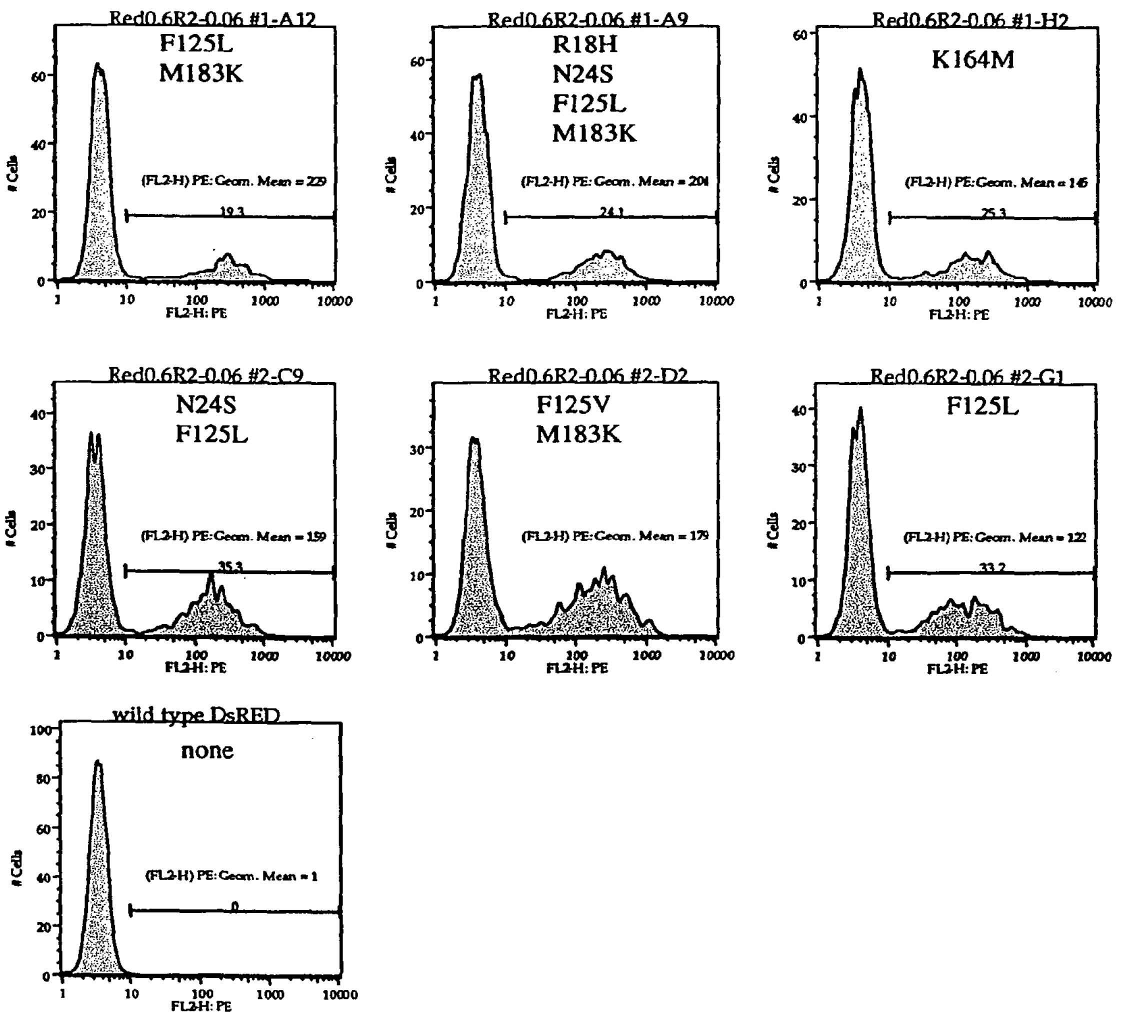
<u>DNA and translated protein sequence of codon optimized wt DsRED protein</u>

1/1 31/11
ATG GTG CGC TCC TCC AAG AAC GTC ATC AAG GAG TTC ATG CGC TTC AAG GTG CGC ATG GAG
M V R S S K N V I K E F M R F K V R M E
61/21 91/31
GGC ACC GTG AAC GGC CAC GAG TTC GAG ATC GAG GGC GAG GGC GAG GGC CGC CCC TAC GAG
G T V N G H E F E I E G E G E G R P Y E
121/41 151/51
GGC CAC AAC ACC GTG AAG CTG AAG GTG ACC AAG GGC GGC CCC CTG CCC TTC GCC TGG GAC
G H N T V K L K V T K G G P L P F A W D
181/61 211/71
ATC CTG TCC CCC CAG TTC CAG TAC GGC TCC AAG GTG TAC GTG AAG CAC CCC GCC GAC ATC
I L S P Q F Q Y G S K V Y V K H P A D I
241/81 271/91
CCC GAC TAC AAG AAG CTG TCC TTC CCC GAG GGC TTC AAG TGG GAG CGC GTG ATG AAC TTC
P D Y K K L S F P E G F K W E R V M N F
301/101 331/111
GAG GAC GGC GGC GTG GTG ACC GTG ACC CAG GAC TCC TCC CTG CAG GAC GGC TGC TTC ATC
E D G G V V T V T Q D S S L Q D G C F I
361/121 391/131
TAC AAG GTG AAG TTC ATC GGC GTG AAC TTC CCC TCC GAC GGC CCC GTA ATG CAG AAG AAG
Y K V K F I G V N F P S D G P V M Q K K
421/141 451/151
ACC ATG GGC TGG GAG GCC TCC ACC GAG CGC CTG TAC CCC CGC GAC GGC GTG CTG AAG GGC
T M G W E A S T E R L Y P R D G V L K G
481/161 511/171
GAG ATC CAC AAG GCC CTG AAG CTG AAG GAC GGC GGC CAC TAC CTG GTG GAG TTC AAG agt
E I H K A L K L K D G G H Y L V E F K S
541/181 571/191
ATC TAC ATG GCC AAG AAG CCC GTG CAG CTG CCC GGC TAC TAC TAC GTG GAC TCC AAG CTG
I Y M A K K P V Q L P G Y Y Y V D S K L
601/201 631/211
GAC ATC ACC TCC CAC AAC GAG GAC TAC ACC ATC GTG GAG CAG TAC GAG CGC ACC GAG GGC
D I T S H N E D Y T I V E Q Y E R T E G
661/221 691/231
CGC CAC CAC CTG TTC CTG gag gag gcc gcc aag gcc gac tac aag gac gac gac gac aag
R H H L F L E E A A K A D Y K D D D D K
721/241
TAG
•

List of confirmed Mutations from screen:
N24S, N24R, N24H (orange box)
F125L, and F125V (blue box)
K164M (green box)
M183K (red box)

Figure 2. Examples of Brighter DsRED mutants.

Individual mutant sequences (mutations shown in red) were infected into a population of BJAB ells and analyzed by FACS 48hours after infection. Wild type DsREDcFlag is shown in the bottom left and exhibits no FL2 fluorescence above background autofluorescence. Since the infection rate was realtively low, most cells express the DsRED variant from a single integrated retrovirus. All histograms have FL2 (normally measures PE on the FACS) on the x-axis. For reference, GFP is measured in FL1.



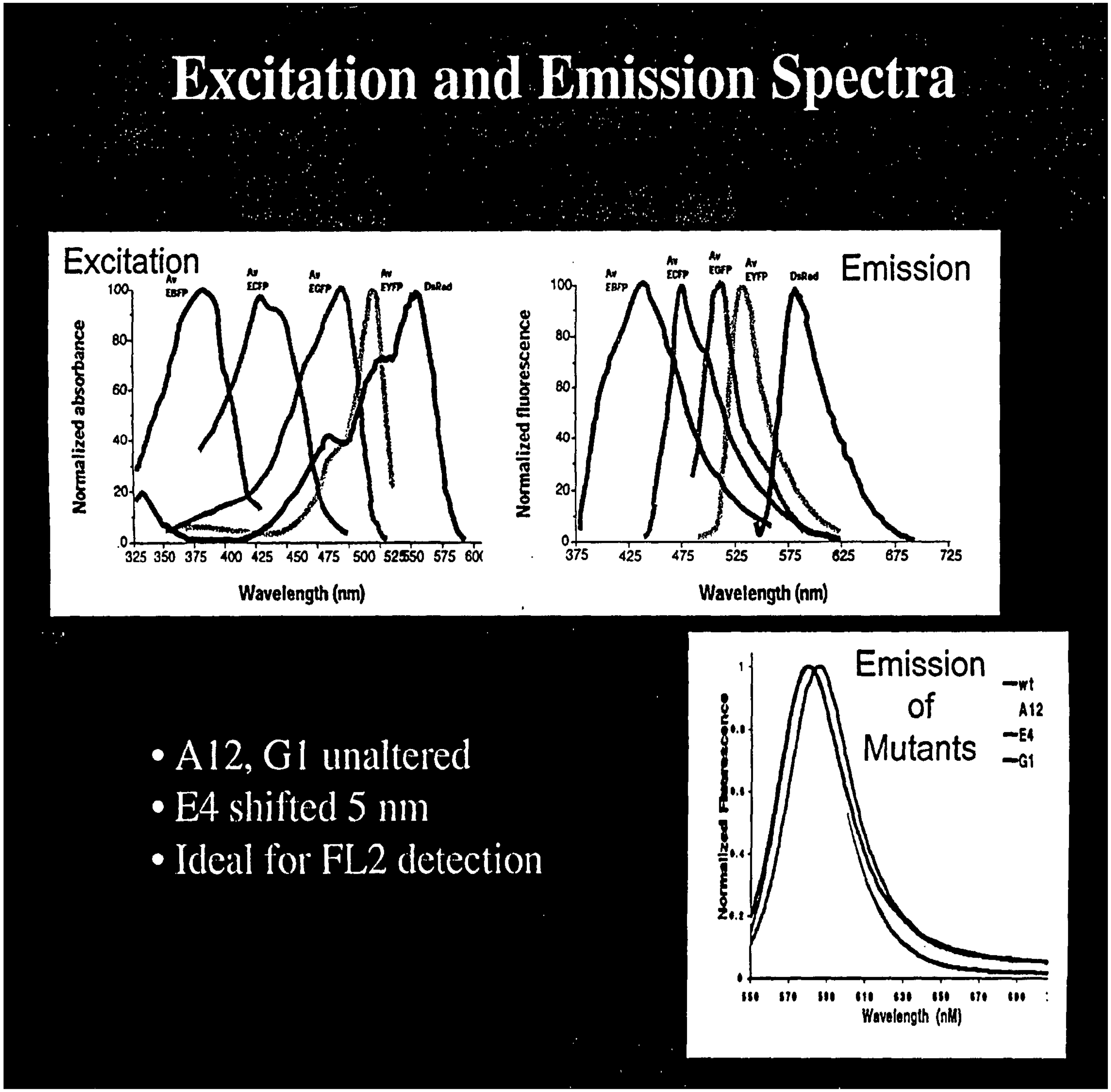
Brightest clones by mean FL2 fluorescence

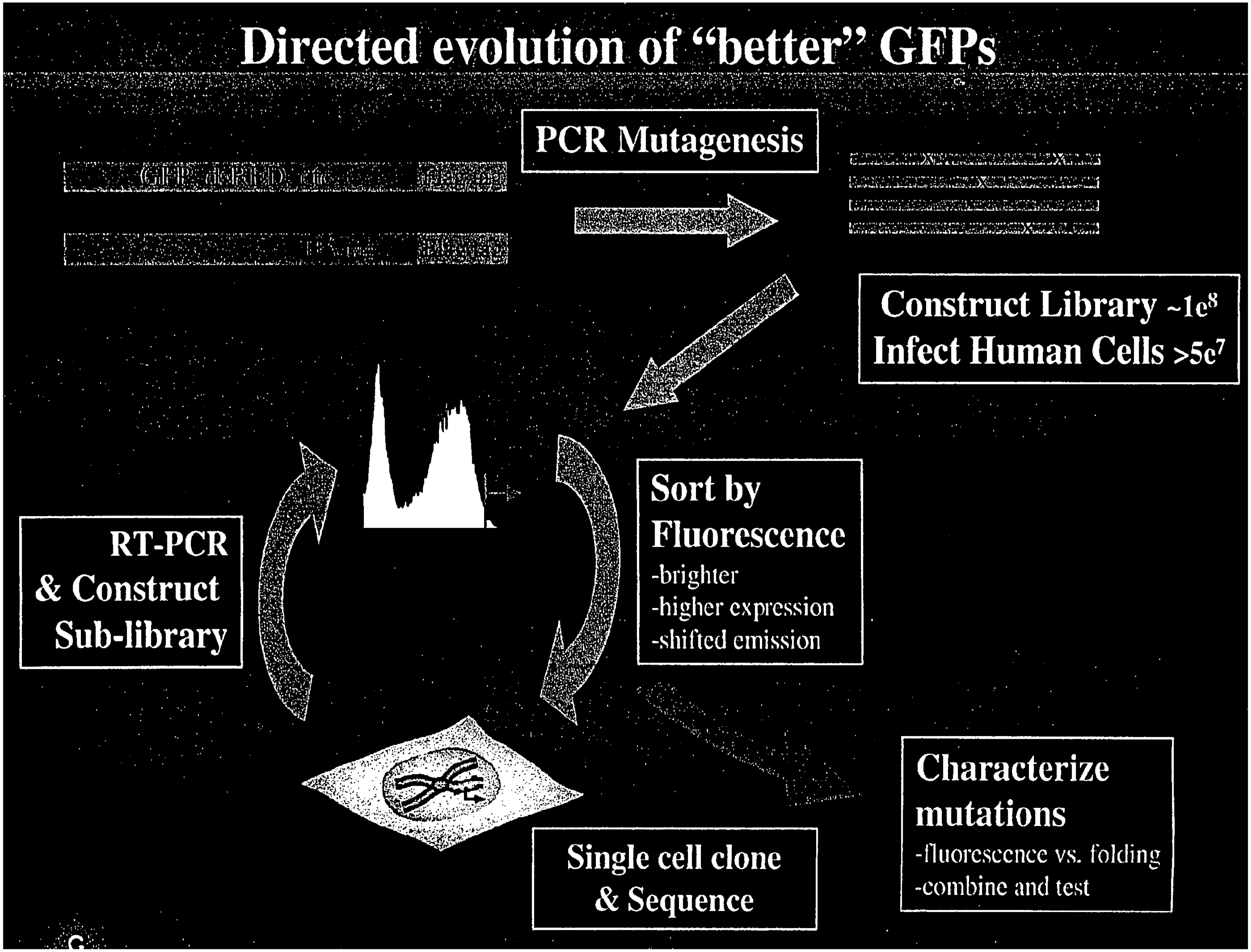
Clone	$m_{G\ (FL2)}$	%FL2+	m_G retest	%FL2+	mutations	identical/48
1-A12	229	16	101	4.2	F124L,M182K +	1
2-A04	189	27.4	98	3.1	F124L,M182K,K138	0
1-G09	186	17.8	97	4.7	F124L,M182K,N23S +	2
1-A09	204	19.7	85	4.5	F124L,M182K,N23S,R18H	1 (-M182K)
2-D02	179	30.1	82	3	F124V,M182	1
1-H02	145	18.6	81	4.5	K163M	1
1-E04	130	21.6	78	3.3	K163M	0
1-G12	118	16.7	61	4.8	F124L, E149D, D170G	0
2-G01	122	21	44	3.6	F124L	13
AsRED		-	17	1.7	wt	-
dsREDcflag	-	-	<2	0	wt	-

by mutation: (out of 48)

46 F124L/V,
10 M182K/T
8 N23S/R/H,
3 K92R

8 F124 + M182
5 N23 + F124


Excitation and Emission Spectra
Excitation
Normalized absorbance
Av EBFP
Av ECFP
Av EGFP
Av EYFP
DsRed
Wavelength (nm)
Emission
Normalized fluorescence
Wavelength (nm)
Emission of Mutants
wt
A12
E4
G1
Wavelength (nM)
• A12, G1 unaltered
• E4 shifted 5 nm
• Ideal for FL2 detection


Directed evolution of "better" GFPs
PCR Mutagenesis
Construct Library ~1e[8]
Infect Human Cells >5e[7]
Sort by
Fluorescence
-brighter
-higher expression
-shifted emission
RT-PCR
& Construct
Sub-library
Single cell clone
& Sequence
Characterize
mutations
-fluorescence vs. folding
-combine and test